

CO-CLUSTERING SEPARATELY EXCHANGEABLE NETWORK DATA

BY DAVID CHOI^{*} AND PATRICK J. WOLFE[†]

Carnegie Mellon University^{} and University College London[†]*

This article establishes the performance of stochastic blockmodels in addressing the co-clustering problem of partitioning a binary array into subsets, assuming only that the data are generated by a nonparametric process satisfying the condition of separate exchangeability. We provide oracle inequalities with rate of convergence $\mathcal{O}_P(n^{-1/4})$ corresponding to profile likelihood maximization and mean-square error minimization, and show that the blockmodel can be interpreted in this setting as an optimal piecewise-constant approximation to the generative nonparametric model. We also show for large sample sizes that detection of co-clusters in such data indicates with high probability the existence of co-clusters of similar proportion and connectivity in the generative process.

1. Introduction. Blockmodels are popular tools for network modeling that see wide and rapidly growing use in analyzing social, economic, and biological systems; see [13, 26] for recent overviews. This article establishes the performance of stochastic blockmodels for the co-clustering problem [15, 24] of partitioning a binary array into subsets, assuming *only that the data are generated by a general nonparametric process* satisfying the condition of separate exchangeability [12]. This significantly generalizes known results for the blockmodel and its co-blockmodel variant, which have only recently been established under the requirement that the model be correctly specified [3, 4, 9, 10, 14, 15, 23, 24, 27].

The stochastic blockmodel provides a natural parametric approximation in the nonparametric setting we consider [4]. We quantify this notion by deriving an oracle inequality which states that the maximum profile likelihood estimate asymptotically minimizes a Kullback–Leibler divergence risk functional linking the class of co-blockmodels and the nonparametric generative process, and also give rates at which it is possible to asymptotically minimize L^2 risk. Additionally, we show that detection of co-clusters by either of these methods implies with high probability the existence of co-clusters of similar proportion and connectivity in the generative process.

AMS 2000 subject classifications: Primary: 62G05; secondary: 05C80, 60B20

Keywords and phrases: bipartite graph, network clustering, oracle inequality, profile likelihood, statistical network analysis, stochastic blockmodel and co-blockmodel

We also show that the co-blockmodel can identify extremal clusterings in data—*network communities*—even if the actual generative process is far from a blockmodel. Our results thus motivate a variety of practical algorithms for statistical network analysis in the area of *community detection*. A great deal of attention and effort has been devoted to this task [13, 16, 22, 26], but the qualitative interpretability of results remains greatly hampered by the fact that the validity of any chosen model in correctly specifying cluster-like behavior is often highly debatable.

Our results imply that community detection can instead be understood quantitatively as finding a best piecewise-constant or simple function approximation to a flexible nonparametric process. In settings where the underlying generative process is not well understood and the specification of highly structured models is thus premature, such an approach is natural for exploratory data analysis. Such usage of blockmodel estimates has even been likened to—and may even come to be as canonically accepted as—the use of histograms to characterize exchangeable data in non-network settings [3].

The article is organized as follows. In Section 2, we introduce our nonparametric setting and the stochastic co-blockmodel. In Section 3 we present oracle inequalities for co-clustering based on blockmodel fitting. In Section 4 we derive a general consistency result, showing that the collection of extremal co-clusterings of the data converges to that of a generative nonparametric process. We prove this result in Section 5, by combining a construction used to establish a theory of graph limits [5, 6, 7] with statistical learning theory results on U-statistics [11]. In Section 6 we conclude with a brief discussion of our results and their relation to other recent papers, such as [9] and [15]. Appendices A and B contain additional proofs and technical lemmas.

2. Model elicitation. Denote by $G = (V_1, V_2, E)$ an observed bipartite graph with edge set E and vertex sets (V_1, V_2) , where we assume known assignment of vertices to V_1 or V_2 . For example, V_1 and V_2 might respectively represent people and locations, with edge (i, j) denoting that person i frequents location j . Let A be the adjacency matrix of G .

2.1. Exchangeable graph models. Exchangeability implies that the node ordering of a graph carries no information [3, 18]. For a bipartite graph represented as a binary array A , the appropriate notion is as follows.

DEFINITION 2.1 (Separate exchangeability [12]). *Let $\{A_{ij}\}_{i,j=1}^\infty$ be binary random variables. They are said to be separately exchangeable if*

$$P(A_{ij} = X_{ij}, 1 \leq i, j \leq n) = P(A_{ij} = X_{\Pi_1(i)\Pi_2(j)}, 1 \leq i, j \leq n)$$

for all $n = 1, 2, \dots$, all permutations Π_1, Π_2 of $1, \dots, n$, and all $X \in \{0, 1\}^{n \times n}$.

If we identify a finite set of rows and columns of A with the adjacency matrix of a bipartite graph, then it is clear that the notion of separate exchangeability encompasses a broad class of network models. Indeed, given a single sample of an unlabeled graph, it is natural to consider models that are invariant to permutation of its adjacency matrix; see [3, 18] for discussion.

The assumption of separate exchangeability is the only one we will require for our results to hold. A representation of models in this class is given by the Aldous–Hoover theorem for separately exchangeable binary arrays.

DEFINITION 2.2 (Exchangeable array model). *Fix a measurable mapping $\omega : [0, 1]^3 \rightarrow [0, 1]$. Then the following model generates an exchangeable random bipartite graph $G = (V_1, V_2, E)$ through its adjacency matrix A .*

1. Generate $\alpha \sim \text{Uniform}(0, 1)$;
2. Fix $m = |V_1|$ and $n = |V_2|$, and generate each element of $\xi = (\xi_1, \dots, \xi_m)$ and $\zeta = (\zeta_1, \dots, \zeta_n) \stackrel{iid}{\sim} \text{Uniform}(0, 1)$;
3. For $i = 1, \dots, m$, and $j = 1, \dots, n$, generate $A_{ij} \stackrel{iid}{\sim} \text{Bernoulli}(\omega^\alpha(\xi_i, \zeta_j))$, where $\omega(x, y) \equiv \omega^\alpha(x, y)$ denotes the function $(x, y) \mapsto \omega(\alpha, x, y)$. If $A_{ij} = 1$, then connect vertices $i \in V_1$ and $j \in V_2$.

The Aldous–Hoover theorem states that this representation is sufficient to describe any separately exchangeable network distribution.

THEOREM 2.1 (Aldous–Hoover [12]). *Let $\{A_{ij}\}_{i,j=1}^\infty$ be a separately exchangeable binary array. Then there exists some $\omega : [0, 1]^3 \rightarrow [0, 1]$, unique up to measure-preserving transformation, which generates $\{A_{ij}\}_{i,j=1}^\infty$.*

The interpretation of the exchangeable graph model of Definition 2.2 is that each vertex has a latent parameter in $[0, 1]$ (ξ_i for vertex i in V_1 , and ζ_j for vertex j in V_2) which determines its affinity for connecting to other vertices, while α is a network-wide connectivity parameter (non-identifiable from a single network sample). Because ξ and ζ are latent, $\omega(x, y)$ itself is identifiable only up to measure-preserving transformation, and is hence indistinguishable from any mapping $(x, y) \mapsto \omega(\alpha, \pi_1(x), \pi_2(y))$ for which π_1, π_2 are in the set \mathcal{P} of measure-preserving maps on $[0, 1]$. We will identify members of this equivalence class in the sequel.

2.2. The stochastic co-blockmodel. Many popular network models can be recognized as instances of Definition 2.2. For example, [1, 19, 20] all present

models in which the resulting $\omega(\alpha, x, y)$ is constant in α , while [21] requires the full parameterization $\omega(\alpha, x, y)$. The stochastic co-blockmodel specifies $\omega(\alpha, x, y)$ constant in α and also piecewise-constant in x and y , and thus can be viewed as a simple function approximation to $\omega(x, y)$ in Definition 2.2.

DEFINITION 2.3 (Stochastic co-blockmodel). *Fix integers $K_1, K_2 > 0$, a matrix $\theta \in [0, 1]^{K_1 \times K_2}$, and discrete probability distributions μ and ν over $1, \dots, K_1$ and $1, \dots, K_2$. Then the stochastic co-blockmodel generates an exchangeable bipartite graph $G = (V_1, V_2, E)$ through the matrix A as follows.*

1. *Fix $m = |V_1|$ and $n = |V_2|$, and generate $S = (S(1), \dots, S(m)) \stackrel{iid}{\sim} \mu$ and $T = (T(1), \dots, T(n)) \stackrel{iid}{\sim} \nu$;*
2. *For $i = 1, \dots, m$, and $j = 1, \dots, n$, generate $A_{ij} \stackrel{iid}{\sim} \text{Bernoulli}(\theta_{S(i)T(j)})$. If $A_{ij} = 1$, then connect vertices $i \in V_1$ and $j \in V_2$.*

Additionally, given co-blockmodel parameters $\phi \equiv (\mu, \nu, \theta)$, define

$$\omega_\phi(x, y) = \theta_{F_\mu^{-1}(x)F_\nu^{-1}(y)}, \quad x, y \in [0, 1],$$

as the mapping corresponding to Definition 2.2, with $F_\mu^{-1}(x) = \inf_z \{F_\mu(z) \geq x\}$ the inverse distribution function corresponding to a given distribution μ .

Without loss of generality we assume $K_1 = K_2 = K$ in the sequel, noting that our results do not depend in any crucial way on this assumption. Thus, vertices in V_1 belong to one of K latent classes, as do those in V_2 . The matrix $\theta \in [0, 1]^{K \times K}$ indexes the corresponding connection affinities between classes in V_1 and V_2 . As S and T are assumed latent, the stochastic co-blockmodel is identifiable only up to a permutation of class labels.

3. Oracle inequalities for co-clustering. If we assume that the separately exchangeable data model of Definition 2.2 is in force, then it is natural to approximate $\omega(x, y)$ by way of $\omega_\phi(x, y)$ according to the stochastic co-blockmodel of Definition 2.3. This approximation task is equivalent to fixing K and estimating $\phi = (\mu, \nu, \theta)$ by co-clustering the entries of an observed adjacency matrix $A \in \{0, 1\}^{m \times n}$ corresponding to a bipartite network.

To accomplish this task, we consider M-estimators that involve an optimization over latent blockmodel variables S and T . To describe these estimators, we must consider the set of all possible co-clusterings of A . To this end, let the set Φ contain all (μ, ν, θ) of the form $(\mu, \nu, \theta) \in \Omega_m \times \Omega_n \times [0, 1]^{K \times K}$, where Ω_m denotes the set of all probability distributions over $\{1, \dots, K\}$ whose elements are integer multiples of $1/m$:

$$\Omega_m = \left\{ p \in \left\{ 0, \frac{1}{m}, \frac{2}{m}, \dots, 1 \right\}^K : \sum_{a=1}^K p_a = 1 \right\},$$

and let \mathcal{Q}_μ^m denote the set of all assignment functions that partition the set $\{1, \dots, m\}$ into K classes in a manner that respects the proportions dictated by $\mu = (\mu_1, \dots, \mu_K) \in \Omega_m$:

$$\mathcal{Q}_\mu^m = \{v \in \{1, \dots, K\}^m : |v^{-1}(a)| = m\mu_a, a = 1, \dots, K\}.$$

Note that by construction, any estimator $\hat{\phi}(A) = (\hat{\mu}, \hat{\nu}, \hat{\theta})$ based on an empirical co-clustering of the observed data $A \in \{0, 1\}^{m \times n}$ has codomain Φ .

We now establish that, for L^2 risk and Kullback–Leibler divergence, there exist M-estimators that enable us to determine, with rate of convergence $n^{-1/4}$, optimal piecewise-constant approximations of the generative $\omega(x, y)$, up to quantization due to the discreteness of Φ .

THEOREM 3.1 (Oracle inequalities for co-clustering). *Let $A \in \{0, 1\}^{m \times n}$ be a separately exchangeable array generated by some ω in accordance with Definition 2.2, and consider fitting a K -class stochastic co-blockmodel parameterized by $\phi \equiv (\mu, \nu, \theta)$ to A . Then as $n \rightarrow \infty$, with K and m/n fixed,*

1. *For the co-blockmodel M-estimator*

$$(3.1) \quad \hat{\phi} = \underset{\phi \in \Phi}{\operatorname{argmin}} \left\{ \min_{S \in \mathcal{Q}_\mu^m, T \in \mathcal{Q}_\nu^n} \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n |\theta_{S(i)T(j)} - A_{ij}|^2 \right\}$$

relative to the L^2 risk

$$R_\omega(\phi) = \inf_{\pi_1, \pi_2 \in \mathcal{P}} \int_{[0,1]^2} |\omega(\pi_1(x), \pi_2(y)) - \omega_\phi(x, y)|^2 dx dy,$$

we have that

$$R_\omega(\hat{\phi}) - \inf_{\phi \in \Phi} R_\omega(\phi) = \mathcal{O}_P(n^{-1/4});$$

2. *For the profile likelihood co-blockmodel M-estimator*

$$(3.2) \quad \hat{\phi} = \underset{\phi \in \Phi}{\operatorname{argmax}} \left\{ \max_{S \in \mathcal{Q}_\mu^m, T \in \mathcal{Q}_\nu^n} \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \left\{ A_{ij} \log(\theta_{S(i)T(j)}) \right. \right. \\ \left. \left. + (1 - A_{ij}) \log(1 - \theta_{S(i)T(j)}) \right\} \right\}$$

relative to

$$L_\omega(\phi) = \sup_{\pi_1, \pi_2 \in \mathcal{P}} \int_{[0,1]^2} \left\{ \omega(\pi_1(x), \pi_2(y)) \log \omega_\phi(x, y) \right. \\ \left. + [1 - \omega(\pi_1(x), \pi_2(y))] \log(1 - \omega_\phi(x, y)) \right\} dx dy,$$

we have whenever $\operatorname{argmax}_{\phi \in \Phi} L_\omega(\phi)$ exists that

$$(3.3) \quad \frac{\max_{\phi \in \Phi} L_\omega(\phi) - L_\omega(\hat{\phi})}{B(\operatorname{argmax}_{\phi \in \Phi} L_\omega(\phi)) + B(\hat{\phi})} = \mathcal{O}_P(n^{-1/4}),$$

with $B(\phi) = B(\theta(\phi)) = \max_{1 \leq a, b \leq K} |\log(\theta_{ab}/(1 - \theta_{ab}))|$.

Theorem 3.1 is proved in Appendix A. Its first result establishes that minimization of the squared error between a fitted co-blockmodel and the data according to (3.1) serves as a proxy for approximation of ω by ω_ϕ in mean square, while its second result establishes that fitting a stochastic co-blockmodel via profile likelihood according to (3.2) is equivalent to minimizing the average Kullback–Leibler divergence of the approximation $\omega_\phi(x, y)$ from the generative $\omega(x, y)$.

While the necessary optimizations in (3.1) and (3.2) are not currently known to admit efficient exact algorithms, they strongly resemble existing objective functions for community detection for which many authors have reported good heuristics [16, 22, 26]. Furthermore, polynomial-time spectral algorithms are known in certain settings to find correct labelings under the assumption of a generative blockmodel [14, 23], suggesting that efficient algorithms may exist when distinct clusterings or community divisions are present in the data. In this vein, a universal thresholding procedure based on the singular value decomposition has been very recently proposed in [9].

REMARK 3.1. *The objective function of (3.2) can be replaced by the full profile likelihood*

$$\begin{aligned} \max_{S \in \mathcal{Q}_\mu^m, T \in \mathcal{Q}_\nu^n} & \left\{ \sum_{i=1}^m \log \mu_{S(i)} + \sum_{j=1}^n \log \nu_{T(j)} \right. \\ & \left. + \sum_{i=1}^m \sum_{j=1}^n \{A_{ij} \log \theta_{S(i)T(j)} + (1 - A_{ij}) \log(1 - \theta_{S(i)T(j)})\} \right\}, \end{aligned}$$

and the same rate of convergence can then be established with respect to the corresponding term for $L_\omega(\phi)$, adapting the proofs in Appendices A and B.

REMARK 3.2. *Let $\bar{\phi} = \operatorname{argmax}_{\phi \in \Phi} L_\omega(\phi)$. Terms $B(\bar{\phi})$ and $B(\hat{\phi})$ in (3.3) indicate that elements of $\bar{\theta}$ and $\hat{\theta}$ must be bounded away from zero and one as $n \rightarrow \infty$; otherwise $L_\omega(\hat{\phi})$ can be much smaller than $L_\omega(\bar{\phi})$.*

This is a natural consequence of the fact that the Kullback–Leibler divergence of ω_ϕ from ω is finite if and only if ω is absolutely continuous with respect to ω_ϕ . To see that this must be the case, consider ξ, ζ , and A generated according to the model of Definition 2.2 with ω constant in α as

$$\omega(x, y) = \begin{cases} 1 & \text{if } x \leq 1/2, y \leq 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mu_1 = m^{-1} \sum_{i=1}^m 1\{\xi_i \leq 1/2\}$, and let $\nu_1 = n^{-1} \sum_{j=1}^n 1\{\zeta_j \leq 1/2\}$. It can be seen that $\omega_{\hat{\phi}}$ corresponding to the maximum-likelihood blockmodel fit is

$$\omega_{\hat{\phi}}(x, y) = \begin{cases} 1 & \text{if } x \leq \mu_1, y \leq \nu_1, \\ 0 & \text{otherwise.} \end{cases}$$

Thus $L_\omega(\hat{\phi}) = -\infty$ unless $\mu_1 = \nu_1 = 1/2$.

4. Convergence of extremal co-clusterings. Theorem 3.1 provides oracle inequalities which state that a co-blockmodel fitted to separately exchangeable network data will be near-optimal in terms of minimizing risk. We now show that the fitted model is interpretable, in the sense that co-clusters of similar proportion and connectivity to those fitted will exist with high probability in the generative process of Definition 2.2.

Our next theorem can also be seen as a general consistency result in its own right, and is the main technical tool necessary to obtain the rate of convergence $\mathcal{O}_P(n^{-1/4})$ in Theorem 3.1. It states that the collection of extremal co-clusterings of the data converges to the collection of extremal co-clusterings of the generative nonparametric process of Definition 2.2.

As in Section 3, we require a means of indexing the set of all possible co-clusterings of any bipartite adjacency matrix A and nonparametric generating function $\omega(x, y)$. To this end, for distributions μ, ν , adjacency matrix $A \in \{0, 1\}^{m \times n}$, and $\omega : [0, 1]^2 \rightarrow [0, 1]$ as in Sections 2 and 3, we will define sets of matrices $\mathcal{F}_{\mu\nu}^A$ and $\mathcal{F}_{\mu\nu}^\omega$ to represent the set of possible co-clusters that can be induced, respectively from the data and from the generative process, by all partitions with class proportions specified by μ and ν .

Our main result is that the convex hulls of $\mathcal{F}_{\mu\nu}^A$ and $\mathcal{F}_{\mu\nu}^\omega$ converge, implying convergence of minimum-risk estimates for any risk functional that achieves its minimum for fixed μ and ν at an extremal point of the convex hull of the feasible set. To make this notion precise, we first require the following.

4.1. Representing sets of co-clusterings and their extrema. Given a bipartite graph $G = (V_1, V_2, E)$ with adjacency matrix $A \in \{0, 1\}^{m \times n}$, let

$S \in \{1, \dots, K\}^m$ and $T \in \{1, \dots, K\}^n$ partition V_1 and V_2 respectively into K subsets. Let $A/ST \in [0, 1]^{K \times K}$ count the number of edges spanning each subset pair, normalized by total number mn of possible edges:

$$(A/ST)(a, b) = \frac{1}{mn} \sum_{i \in S^{-1}(a)} \sum_{j \in T^{-1}(b)} A_{ij}, \quad a, b = 1, \dots, K.$$

Similarly, given $\omega : [0, 1]^2 \rightarrow [0, 1]$ and mappings $\sigma, \tau : [0, 1] \rightarrow \{1, \dots, K\}$, let $\omega/\sigma\tau \in [0, 1]^{K \times K}$ encode the mass of ω assigned to each subset pair.

$$(\omega/\sigma\tau)(a, b) = \int_{\sigma^{-1}(a) \times \tau^{-1}(b)} \omega(x, y) dx dy, \quad a, b = 1, \dots, K.$$

Let \mathcal{Q}_μ^m be defined as in Section 3, and let \mathcal{Q}_μ analogously denote the set of partitions of $[0, 1]$ into K subsets whose cardinalities are of proportions μ_1, \dots, μ_K :

$$\mathcal{Q}_\mu = \{\sigma : [0, 1] \rightarrow \{1, \dots, K\} \text{ such that } |\sigma^{-1}(a)| = \mu_a, a = 1, \dots, K\}.$$

We are now ready to state our required definitions.

DEFINITION 4.1 (Sets $\mathcal{F}_{\mu\nu}^A$ and $\mathcal{F}_{\mu\nu}^\omega$ of possible co-clusters). *For fixed discrete probability distributions μ and ν over $1, \dots, K$, we define the sets $\mathcal{F}_{\mu\nu}^A$ and $\mathcal{F}_{\mu\nu}^\omega$ of all co-clustering matrices A/ST and mappings $\omega/\sigma\tau$ induced by $(S, T) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n$ and $(\sigma, \tau) \in \mathcal{Q}_\mu \times \mathcal{Q}_\nu$ as follows:*

$$\begin{aligned} \mathcal{F}_{\mu\nu}^A &= \{A/ST : S \in \mathcal{Q}_\mu^m, T \in \mathcal{Q}_\nu^n\}, \\ \mathcal{F}_{\mu\nu}^\omega &= \{\omega/\sigma\tau : \sigma \in \mathcal{Q}_\mu, \tau \in \mathcal{Q}_\nu\}. \end{aligned}$$

DEFINITION 4.2 (Support function). *Let \mathcal{F} be a non-empty subset of $\mathbb{R}^{K \times K}$, endowed with the standard inner product $\langle \Gamma, F \rangle = \text{tr}(\Gamma^T F)$. We define the support function $h_{\mathcal{F}} : \mathbb{R}^{K \times K} \rightarrow \mathbb{R} \cup \{+\infty\}$ of $\mathcal{F} \subset \mathbb{R}^{K \times K}$ as*

$$h_{\mathcal{F}}(\Gamma) = \sup_{F \in \mathcal{F}} \langle \Gamma, F \rangle.$$

4.2. A general result on consistency of co-clustering. The sets $\mathcal{F}_{\mu\nu}^A, \mathcal{F}_{\mu\nu}^\omega \subset [0, 1]^{K \times K}$ describe all possible co-clustering that can be induced respectively from A and ω with respect to μ and ν . The support function $h_{\mathcal{F}}(\Gamma)$ defines the supporting hyperplanes of such an \mathcal{F} in any direction Γ , thereby providing a representation of its closed convex hull. From the above we see

$$(4.1a) \quad h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) = \max_{(S, T) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \langle \Gamma, A/ST \rangle,$$

$$(4.1b) \quad h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma) = \sup_{(\sigma, \tau) \in \mathcal{Q}_\mu \times \mathcal{Q}_\nu} \langle \Gamma, \omega/\sigma\tau \rangle.$$

Equipped with these notions, we are now ready to state our main result.

THEOREM 4.1. *Let $A \in \{0, 1\}^{m \times n}$ be a separately exchangeable array generated by some ω in accordance with Definition 2.2, and consider fitting a K -class stochastic co-blockmodel to A . Then for each K and each ratio m/n , there exists a universal constant C such that as $n \rightarrow \infty$,*

$$\mathbb{P} \left(\max_{(\mu, \nu) \in \Omega_m \times \Omega_n} \left\{ \sup_{\Gamma \in [-1, 1]^{K \times K}} |h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)| \right\} \geq \frac{C}{n^{1/4}} \right) = o(1),$$

The geometric implication of Theorem 4.1 is as follows.

COROLLARY 4.1. *Under the assumptions of Theorem 4.1, the convex hulls of $\mathcal{F}_{\mu\nu}^A$ and $\mathcal{F}_{\mu\nu}^\omega$ converge in the Hausdorff metric at rate $\mathcal{O}_P(n^{-1/4})$.*

PROOF OF COROLLARY 4.1. Denote the Frobenius or Hilbert-Schmidt metric on $\mathbb{R}^{K \times K}$ induced by $\langle \cdot, \cdot \rangle$ as $d(F, F') = (\sum_{a,b} |F_{ab} - F'_{ab}|^2)^{1/2}$. The Hausdorff distance $d_{\text{Haus}}(\cdot, \cdot)$ between sets $\mathcal{F}, \mathcal{F}' \in \mathbb{R}^{K \times K}$ is then given by

$$d_{\text{Haus}}(\mathcal{F}, \mathcal{F}') = \max \left\{ \sup_{F \in \mathcal{F}} \left\{ \inf_{F' \in \mathcal{F}'} d(F, F') \right\}, \sup_{F' \in \mathcal{F}'} \left\{ \inf_{F \in \mathcal{F}} d(F, F') \right\} \right\}.$$

Recall that it measures the maximal shortest distance d between any two elements of \mathcal{F} and \mathcal{F}' . Given non-empty, totally bounded $\mathcal{F}, \mathcal{F}' \subset \mathbb{R}^{K \times K}$, it holds, with conv denoting the convex hull and $\|\cdot\|$ the Frobenius norm, that

$$d_{\text{Haus}}(\text{conv}(\mathcal{F}), \text{conv}(\mathcal{F}')) = \sup_{\|\Gamma\|=1} |h_{\mathcal{F}}(\Gamma) - h_{\mathcal{F}'}(\Gamma)|,$$

(see, e.g., [25], as applied to the convex hulls of the closures of \mathcal{F} and of \mathcal{F}'). Thus by Theorem 4.1,

$$\max_{(\mu, \nu) \in \Omega_m \times \Omega_n} d_{\text{Haus}}(\text{conv}(\mathcal{F}_{\mu\nu}^A), \text{conv}(\mathcal{F}_{\mu\nu}^\omega)) = \mathcal{O}_P(n^{-1/4}).$$

□

The result of Theorem 4.1 can be directly related to work in [6, 7], a pair of articles which explores in depth the notion of a graph limit as n goes to infinity, and the statistical applications thereof as discussed in [5]. Very broadly speaking, [6, Theorem 2.9] and [7, Theorem 4.6] analyze sets which resemble $\cup_{\mu, \nu} \mathcal{F}_{\mu\nu}^A$ and $\cup_{\mu, \nu} \mathcal{F}_{\mu\nu}^\omega$, and are termed quotients. For these sets, the authors show convergence in the Hausdorff metric at rate $\mathcal{O}((\log(n))^{-1/2})$ for a distance d_\square known as the cut metric, and detail many implications thereof. By fixing μ and ν , as required by our M-estimates, we are studying what those authors term the microcanonical quotients. By restricting convergence to the convex hull in Theorem 4.1, as allowed by our M-estimates, we gain an exponentially faster bound on the rate of convergence.

5. Proof of Theorem 4.1. Our proof strategy is inspired by [6] and adapts certain of its tools, but requires $\mathcal{F}_{\mu\nu}^A$ to be studied separately for each choice of μ and ν in order to yield Corollary 4.1 and the oracle inequalities of Theorem 3.1. Most significantly, we do not use the Szemerédi regularity lemma, which typically features strongly in the graph-theoretic literature, and provides a means of partitioning any large dense graph into a small number of regular clusters. Results in this direction are possible, but instead we use a Rademacher complexity bound for U-statistics adapted from [11], allowing us to achieve the improved rates of convergence described above.

5.1. Establishing pointwise convergence. The main step in proving Theorem 4.1 is to establish pointwise convergence of $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma)$ to $h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$ for any fixed Γ . We do this through Proposition 5.1 below, after which we may apply it to a union bound over a covering of all $\Gamma \in [-1, 1]^{K \times K}$ to deduce the result of Theorem 4.1. Appendix B provides a formal statement and proof of this argument, along with proofs of all supporting lemmas.

PROPOSITION 5.1 (Pointwise convergence of $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma)$ to $h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$). *Assume the setting of Theorem 4.1, fixing $m = pn$. Then there exist constants C_K, n_K such that, given any $\Gamma \in [-1, 1]^{K \times K}$, μ, ν, ω , and $A \in \{0, 1\}^{m \times n}$ generated from ω , it holds for all $n \geq n_K$ that*

$$\mathbb{P} \left(\left| h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma) \right| \geq \frac{C_K}{n^{1/4}} \right) \leq 2e^{-\sqrt{n}[2\rho/(\rho+1)]} [1 + o(1)].$$

PROOF OF PROPOSITION 5.1. To obtain the claimed result, we must establish lower and upper bounds on the support function $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma)$ that show its convergence to $h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$ at rate $\mathcal{O}_P(n^{-1/4})$. Recalling the definitions of $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma)$ and $h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$ in (4.1), we first require a statement of Lipschitz conditions on $\langle \Gamma, A/ST \rangle$ and $\langle \Gamma, \omega/\sigma\tau \rangle$. Its proof follows by direct inspection.

LEMMA 5.1. *Define for measurable mappings σ, σ' over $[0, 1]$ the metric*

$$d_{\text{Ham}}(\sigma, \sigma') = \int_{[0,1]} 1\{\sigma(x) \neq \sigma'(x)\} dx,$$

and analogously the standard Hamming distance for sequences, with respect to normalized counting measure. Then for any $\Gamma \in [-1, 1]^{K \times K}$ and $A, A' \in [0, 1]^{m \times n}$, with $(S, T, \omega, \sigma, \tau)$ as defined in Section 4.1, we have that

1. $|\langle \Gamma, A/ST \rangle - \langle \Gamma, A/S'T' \rangle| \leq 2[d_{\text{Ham}}(S, S')/m + d_{\text{Ham}}(T, T')/n];$
2. $|\langle \Gamma, \omega/\sigma\tau \rangle - \langle \Gamma, \omega/\sigma'\tau' \rangle| \leq 2[d_{\text{Ham}}(\sigma, \sigma') + d_{\text{Ham}}(\tau, \tau')];$

3. $|\langle \Gamma, A/ST \rangle - \langle \Gamma, A'/ST \rangle| \leq 1/(mn)$ if A, A' differ by a single entry.

In conjunction with McDiarmid's inequality, these Lipschitz conditions yield the following lower bound on $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma)$, proved in Appendix B.1.

LEMMA 5.2 (Lower bound on $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma)$). *Assume the setting of Theorem 4.1. Then there exist constants C'_K, n'_K such that, given any $\Gamma \in [-1, 1]^{K \times K}$, μ, ν, ω , and $A \in \{0, 1\}^{\rho n \times n}$ generated from ω , for all $n \geq n'_K$,*

$$\mathbb{P} \left(h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) \geq \frac{C'_K}{n^{1/4}} \right) \leq 2e^{-\sqrt{n}[2\rho/(\rho+1)]} [1 + o(1)].$$

The upper bound comes by way of Rademacher complexity arguments. The remainder of this section and Appendix B is devoted to its proof.

LEMMA 5.3 (Upper bound on $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma)$). *Assume the setting of Theorem 4.1. Then there exist constants C''_K, n''_K such that, given any $\Gamma \in [-1, 1]^{K \times K}$, μ, ν, ω , and $A \in \{0, 1\}^{\rho n \times n}$ generated from ω, p for all $n \geq n''_K$,*

$$\mathbb{P} \left(h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma) \geq \frac{C''_K}{n^{1/4}} \right) \leq 2e^{-\sqrt{n}[2\rho/(\rho+1)]} [1 + o(1)].$$

Proposition 5.1 now follows simply by combining Lemmas 5.2 and 5.3. \square

5.2. *Establishing an upper bound on $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma)$.* Lemma 5.3 represents the main technical hurdle in obtaining the polynomial rate of convergence given in Theorems 3.1 and 4.1. To illustrate the main ideas as clearly as possible, we will introduce our Rademacher complexity arguments below for the case $K = 2$, deferring the necessary generalizations to Appendix B.

We first define $W \in [0, 1]^{m \times n}$ with reference to Definition 2.2 as

$$W_{ij} = \omega(\xi_i, \zeta_j), \quad i \in 1, \dots, m, \quad j \in 1, \dots, n;$$

and then define, in direct analogy to $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma)$,

$$h_{\mathcal{F}_{\mu\nu}^W}(\Gamma) = \max_{(S,T) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \langle \Gamma, W/ST \rangle = \max_{(S,T) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \left\{ \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n W_{ij} \Gamma_{S(i)T(j)} \right\}.$$

The matrix W serves as an empirical realization of the mapping ω , with its support function $h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)$ defined with respect to co-blockmodel partitions $(S, T) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n$. As proved in Appendix B.2, Lemma 5.4 enables us to bound $|h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - \mathbb{E} h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)|$ using the Lipschitz conditions in Lemma 5.1.

LEMMA 5.4. *Fix some measurable $\omega : [0, 1]^2 \rightarrow [0, 1]$, with $W \in [0, 1]^{m \times n}$ generated by ω and $A \in \{0, 1\}^{m \times n}$ generated by W , and some $\Gamma \in [-1, 1]^{K \times K}$. Then for any $\epsilon > 0$,*

$$(5.1) \quad \mathbb{P} \left(|h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - \mathbb{E} h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)| \geq 2\epsilon \right) \leq 2e^{-2m\epsilon^2/(m+n)} + 2K^{m+n}e^{-2m\epsilon^2}.$$

Having bounded $|h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - \mathbb{E} h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)|$, we must upper-bound $\mathbb{E} h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)$ in terms of $h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$. We will do this in a series of steps, first bounding $\mathbb{E} h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)$ using a result adapted from [2] and proved in Appendix B.3.

LEMMA 5.5. *Let \mathcal{I} and \mathcal{J} be sets of deterministic size, whose elements are sampled without replacement from $1, \dots, m$ and $1, \dots, n$. Let W be generated as in Lemma 5.4, and fix $\Gamma \in [-1, 1]^{K \times K}$. Given $W, \mathcal{I}, \mathcal{J}$, and $(Q, R) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n$, let $\hat{S}^R \equiv \hat{S}^{R, \mathcal{J}, W}$ and $\hat{T}^Q \equiv \hat{T}^{Q, \mathcal{I}, W}$ denote partitions satisfying*

$$(5.2) \quad \hat{S}^R \in \operatorname{argmax}_{S \in \mathcal{Q}_\mu^m} \left\{ \sum_{i=1}^m \sum_{j \in \mathcal{J}} W_{ij} \Gamma_{S(i)R(j)} \right\},$$

$$(5.3) \quad \hat{T}^Q \in \operatorname{argmax}_{T \in \mathcal{Q}_\nu^n} \left\{ \sum_{i \in \mathcal{I}} \sum_{j=1}^n W_{ij} \Gamma_{Q(i)T(j)} \right\}.$$

Then

$$(5.4) \quad \mathbb{E} h_{\mathcal{F}_{\mu\nu}^W}(\Gamma) \leq \mathbb{E} \left(\max_{(Q, R) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \langle \Gamma, W / \hat{S}^R \hat{T}^Q \rangle \right) + K\sqrt{2\pi} \left(|\mathcal{I}|^{-\frac{1}{2}} + |\mathcal{J}|^{-\frac{1}{2}} \right).$$

To bound the right-hand side of (5.4) relative to $h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$, we will introduce an additional construction comprising several steps. Specifically, for fixed (Q, R) and Γ , we will define function classes \mathcal{Q}_U and \mathcal{Q}_V , and a random functional $G_{\sigma\tau}$ which approximates $\langle \Gamma, W / \hat{S}^R \hat{T}^Q \rangle$ for some $(\hat{\sigma}, \hat{\tau}) \in \mathcal{Q}_U \times \mathcal{Q}_V$. By a Rademacher complexity argument, $G_{\hat{\sigma}\hat{\tau}}$ will concentrate for all (Q, R) near its expectation, which itself will be bounded by $h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$.

For the case $K = 2$, define U by

$$U(x) = \sum_{j \in \mathcal{J}} \omega(x, \zeta_j) (\Gamma_{1R(j)} - \Gamma_{2R(j)}).$$

It follows that

$$\hat{S}^R \in \operatorname{argmax}_{S \in \mathcal{Q}_\mu^m} \sum_{i=1}^m U(\xi_i) 1\{S(i) = 1\},$$

and so \hat{S}^R will assign to class 1 the $\mu_1 m$ largest elements of $U(\xi_1), \dots, U(\xi_m)$. If U is invertible, this set can be written $\{\xi_i : U(\xi_i) < t\}$ for some t . To treat

non-invertible U , define \mathcal{Q}_U to be the class of functions $\{1_u : u \in [0, 1]\}$, with 1_u a one-sided interval on the range of U with lexicographic “tie-breaking”:

$$1_u(x) = \begin{cases} 2 & \text{if either } U(x) < U(u), \text{ or } U(x) = U(u) \text{ and } x < u; \\ 1 & \text{if either } U(x) > U(u), \text{ or } U(x) = U(u) \text{ and } x \geq u. \end{cases}$$

Then there exists $\hat{\sigma} \in \mathcal{Q}_U$ such that \hat{S}^R can be chosen to satisfy

$$\hat{S}^R(i) = \hat{\sigma}(\xi_i), \quad i = 1, \dots, m.$$

Let V denote a function defined analogously to U as follows:

$$V(y) = \sum_{i \in \mathcal{I}} \omega(\xi_i, y) (\Gamma_{Q(i)1} - \Gamma_{Q(i)2}),$$

and likewise define \mathcal{Q}_V so that there exists $\hat{\tau} \in \mathcal{Q}_V$ such that \hat{T}^Q can be chosen to satisfy

$$\hat{T}^Q(j) = \hat{\tau}(\zeta_j), \quad j = 1, \dots, n.$$

We are now ready to define $G_{\sigma\tau}$. Given any $\sigma \in \mathcal{Q}_U$ and $\tau \in \mathcal{Q}_V$, let

$$G_{\sigma\tau}(\xi, \zeta) = \frac{1}{mn} \sum_{i \in \overline{\mathcal{I}}} \sum_{j \in \overline{\mathcal{J}}} \omega(\xi_i, \zeta_j) \Gamma_{\sigma(\xi_i)\tau(\zeta_j)},$$

where $\overline{\mathcal{I}}$ is the complement of \mathcal{I} in $\{1, \dots, m\}$, and $\overline{\mathcal{J}}$ the complement of \mathcal{J} in $\{1, \dots, n\}$. Comparing $G_{\sigma\tau}$ to Lemma 5.5, we see that $G_{\hat{\sigma}\hat{\tau}}$ well approximates $\langle \Gamma, W / \hat{S}^R \hat{T}^Q \rangle$ whenever $|\mathcal{I}|$ and $|\mathcal{J}|$ are small; and indeed, we will later set $|\mathcal{I}| = |\mathcal{J}| = n^{1/2}$ in order to obtain an upper bound for $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$.

By construction, the random classes \mathcal{Q}_U and \mathcal{Q}_V are independent of the random variables $\{\xi_i\}_{i \in \overline{\mathcal{I}}}$ and $\{\zeta_j\}_{j \in \overline{\mathcal{J}}}$ appearing in the summand of $G_{\sigma\tau}$. As a result, we may bound the deviation δ_{UV} of $G_{\sigma\tau}$ from its expectation,

$$\delta_{UV} = \sup_{(\sigma, \tau) \in \mathcal{Q}_U \times \mathcal{Q}_V} |G_{\sigma\tau}(\xi, \zeta) - \mathbb{E}(G_{\sigma\tau}(\xi, \zeta) | U, V)|,$$

using Rademacher complexity results for U-statistics [11, Lemma A.1], [17], applied to the class of one-sided interval functions.

LEMMA 5.6. *Assume the setting of Lemma 5.5, and set $\ell = \min(m - |\mathcal{I}|, n - |\mathcal{J}|)$. Then the deviation δ_{UV} of $G_{\sigma\tau}$ from its expectation satisfies*

$$\mathbb{E} \left(\max_{(Q, R) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \delta_{UV} \right) \leq 4 \sqrt{\frac{(|\mathcal{I}| + |\mathcal{J}|) \log K + 2 \binom{K}{2} \log(\ell + 1) + \log 2}{2\ell}}.$$

Lemma 5.6 is proved in Appendix B.5 to hold for arbitrary K , under the appropriate generalization of \mathcal{Q}_U , \mathcal{Q}_V , and quantities that depend on them.

Similarly, we may bound δ_U , defined for $K = 2$ as the maximum discrepancy between the expected and empirical class frequency in \mathcal{Q}_U :

$$\delta_U = \sup_{\sigma \in \mathcal{Q}_U} \left\{ \max_{1 \leq a \leq K} \left| |\sigma^{-1}(a)| - \frac{1}{m} \sum_{i=1}^m 1\{\sigma(\xi_i) = a\} \right| \right\},$$

with δ_V defined mutatis mutandis. We then have the following result, proved for arbitrary K (with appropriate redefinitions of δ_U, δ_V) in Appendix B.6.

LEMMA 5.7. *Assume the setting of Lemma 5.5. Then*

$$\begin{aligned} \mathbb{E} \left(\max_{R \in \mathcal{Q}_\nu^n} \delta_U \right) &\leq 4 \sqrt{\frac{(|\mathcal{J}| + 1) \log K + \binom{K}{2} \log(m + 1) + \log 2}{2m}}, \\ \mathbb{E} \left(\max_{Q \in \mathcal{Q}_\mu^m} \delta_V \right) &\leq 4 \sqrt{\frac{(|\mathcal{I}| + 1) \log K + \binom{K}{2} \log(n + 1) + \log 2}{2n}}. \end{aligned}$$

We state and prove a final auxiliary lemma prior to the proof of Lemma 5.3.

LEMMA 5.8. *Assume the setting of Lemma 5.5. Then*

$$\begin{aligned} \mathbb{E} \left(\max_{(Q,R) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \langle \Gamma, W / \hat{S}^R \hat{T}^Q \rangle \right) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma) &\leq 2 \{ m^{-1} |\mathcal{I}| + n^{-1} |\mathcal{J}| \} \\ &\quad + \mathbb{E} \left(\max_{(Q,R) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \delta_{UV} \right) + 2K \mathbb{E} \left(\max_{(Q,R) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \delta_U + \delta_V \right). \end{aligned}$$

PROOF. Let $\hat{\sigma}$ and $\hat{\tau}$ denote the mappings in \mathcal{Q}_μ and \mathcal{Q}_ν that are respectively closest in the metric d_{Ham} to $\hat{\sigma}$ and $\hat{\tau}$. Observe that we may then expand and upper-bound the left-hand side of the lemma statement by

$$\begin{aligned} &\underbrace{\mathbb{E} \left(\max_{Q,R} \langle \Gamma, W / \hat{S}^R \hat{T}^Q \rangle - G_{\hat{\sigma}\hat{\tau}}(\xi, \zeta) \right)}_{\text{(i)}} + \underbrace{\mathbb{E} \left(\max_{Q,R} G_{\hat{\sigma}\hat{\tau}}(\xi, \zeta) - \langle \Gamma, \omega / \hat{\sigma} \hat{\tau} \rangle \right)}_{\text{(ii)}} \\ &\quad + \underbrace{\mathbb{E} \left(\max_{Q,R} \langle \Gamma, \omega / \hat{\sigma} \hat{\tau} \rangle - \langle \Gamma, \omega / \hat{\sigma} \hat{\tau} \rangle \right)}_{\text{(iii)}} + \underbrace{\mathbb{E} \left(\max_{Q,R} \langle \Gamma, \omega / \hat{\sigma} \hat{\tau} \rangle - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma) \right)}_{\text{(iv)}}, \end{aligned}$$

after which we may upper-bound terms (i)–(iv) in turn as follows.

First, since $|\omega(x, y)\Gamma_{\hat{\sigma}(x)\hat{\tau}(y)}| \leq 1$ for all (x, y) , it follows from their respective definitions that $\langle \Gamma, W/\hat{S}^R \hat{T}^Q \rangle - G_{\hat{\sigma}\hat{\tau}}(\xi, \zeta)$ is deterministically bounded above by $|\mathcal{I}|/m + |\mathcal{J}|/n$. Hence, term (i) is bounded by the same quantity.

Second, observe that by definition, $G_{\hat{\sigma}\hat{\tau}}(\xi, \zeta) - \mathbb{E}(G_{\hat{\sigma}\hat{\tau}}(\xi, \zeta) | U, V) \leq \delta_{UV}$. Since for fixed σ, τ we have $\mathbb{E}(G_{\sigma\tau}(\xi, \zeta) | U, V) = [|\bar{\mathcal{I}}||\bar{\mathcal{J}}|/(mn)] \langle \Gamma, \omega/\sigma\tau \rangle$, with $|\langle \Gamma, \omega/\sigma\tau \rangle| \leq 1$, it holds deterministically that $\mathbb{E}(G_{\hat{\sigma}\hat{\tau}}(\xi, \zeta) | U, V) - \langle \Gamma, \omega/\hat{\sigma}\hat{\tau} \rangle \leq |\mathcal{I}|/m + |\mathcal{J}|/n$. Thus term (ii) is bounded above by the quantity $\mathbb{E}(\max_{(Q,R) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \delta_{UV}) + |\mathcal{I}|/m + |\mathcal{J}|/n$.

Third, by the second Lipschitz condition of Lemma 5.1, we have that $\langle \Gamma, \omega/\hat{\sigma}\hat{\tau} \rangle - \langle \Gamma, \omega/\hat{\sigma}\hat{\tau} \rangle \leq 2[d_{\text{Ham}}(\hat{\sigma}, \hat{\sigma}) + d_{\text{Ham}}(\hat{\tau}, \hat{\tau})]$. Observe that

$$d_{\text{Ham}}(\hat{\sigma}, \hat{\sigma}) \leq \sum_{a=1}^K \left| |\hat{\sigma}^{-1}(a)| - \mu_a \right| \leq \sum_{a=1}^K \left| |\hat{\sigma}^{-1}(a)| - \frac{1}{m} \sum_{i=1}^m 1_{\{\hat{\sigma}(\xi_i)=a\}} \right| \leq K\delta_U,$$

where the second inequality holds as $\hat{S}^R \in \mathcal{Q}_\mu^m$. By the same argument for $d_{\text{Ham}}(\hat{\tau}, \hat{\tau})$, we see term (iii) is bounded by $2K \mathbb{E}(\max_{(Q,R) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \delta_U + \delta_V)$.

To conclude, note term (iv) is deterministically upper-bounded by 0. \square

We may now establish the claimed upper bound on $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$.

PROOF OF LEMMA 5.3. Combining the results of Lemmas 5.4–5.8 yields directly that, with probability at least $1 - 2e^{-2mn\epsilon^2/(m+n)} - 2K^{m+n}e^{-2mn\epsilon^2}$,

$$\begin{aligned} h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma) &\leq 2\epsilon + K\sqrt{2\pi}\{|\mathcal{I}|^{-1/2} + |\mathcal{J}|^{-1/2}\} + 2\{m^{-1}|\mathcal{I}| + n^{-1}|\mathcal{J}|\} \\ &+ f(|\mathcal{I}| + |\mathcal{J}|, \ell, 2\binom{K}{2}) + 2K\left\{f(|\mathcal{I}| + 1, n, \binom{K}{2}) + f(|\mathcal{J}| + 1, m, \binom{K}{2})\right\}, \end{aligned}$$

where $f(p, q, r) = 4\{[p \log K + r \log(q+1) + \log 2]/(2q)\}^{1/2}$, and $\ell = \min(m - |\mathcal{I}|, n - |\mathcal{J}|)$ as in Lemma 5.6. Letting $\epsilon = n^{-1/4}$, $|\mathcal{I}| = |\mathcal{J}| = n^{1/2}$, and fixing $m = \rho n$ as assumed in the hypothesis of Lemma 5.3, it follows that for $n \geq 2$,

$$\begin{aligned} h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma) &\leq \frac{2 + 2K(2\pi)^{1/2} + (4\sqrt{2} + 8K)(2\log K)^{1/2}}{n^{1/4}} \\ &+ \frac{4 + 12(K^2 \log(\rho n + 1) + 2)^{1/2}}{n^{1/2}}, \end{aligned}$$

with probability at least $1 - 2e^{-\sqrt{n}[2\rho/(\rho+1)]} - 2K^{(\rho+1)n}e^{-2\rho n^{3/2}}$. Thus we have that $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma)$ is bounded above by $h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma) + \mathcal{O}_P(n^{-1/4})$, as claimed. \square

6. Discussion. In this article we have addressed the case of network *co-clustering*, in which the inference task is to group two sets of network nodes into classes based on their observed relations. Our results significantly generalize known consistency results for the blockmodel and its co-blockmodel variant: they do not require the data to be generated (even approximately) by a co-blockmodel, and they achieve improved rates of convergence relative to results from the graph limits literature, through the use a Rademacher complexity bound for U-statistics adapted from [11]. The assumption of a nonparametric generative model is both more general and more realistic, and to our knowledge Theorems 3.1 and 4.1 are the first for this regime to establish polynomial rates of convergence.

In [11], these Rademacher complexity results are used to derive convergence rates for learning pairwise rankings. This setting is related to ours, but differs in some important ways. In [11], a rule $r : \mathcal{X} \times \mathcal{X} \rightarrow \{-1, +1\}$ is desired such that, given $X, X' \in \mathcal{X}$, r indicates which has the higher rank. In this setting, X and X' can be thought of as covariates describing the two objects for which a relative ranking is desired, and \mathcal{X} represents the space of allowable covariate values. In our network setting, the nonparametric model $\omega : [0, 1]^2 \rightarrow [0, 1]$ is analogous to a ranking rule, with \mathcal{X} taken to be $[0, 1]$. However, X and X' are never observed in the data, and effectively must be imputed up to measure-preserving transformation.

The recent work of [15] analyzes the consistency of co-clustering with model misspecification, but in a rather different setting, with the data matrix A assumed to be real valued, along with a real-valued generalization of the co-blockmodel. This generalization utilizes discrete latent class variables S and T ; conditioned on $S(i)$ and $T(j)$, the distribution of A_{ij} is assumed to have mean $\theta_{S(i)T(j)}$, but may otherwise be arbitrary up to technical conditions, and may be misspecified in the estimator. Under these assumptions, it is shown that the latent classes can be estimated consistently if their number is known. In the case where A is binary, the conditions of [15] are equivalent to assuming a generative co-blockmodel with known number of classes.

Finally, the very recent work of [9] derives a simple and elegant spectral method to consistently estimate the matrix W defined in the proof of Lemma 5.3 in Section 5.2; i.e., the mapping $\omega(x, y)$, evaluated at the values of the latent variables ξ_1, \dots, ξ_m , and ζ_1, \dots, ζ_n . This implies consistency of estimation of ω in the L^2 sense; and while rates of convergence are not given for general ω , they can be established for particular instances, such as under the assumption of a generative blockmodel whose number of classes K is growing with n . Our setting is distinct, in that we desire only the best blockmodel approximation to ω , and so are able to establish L^2 rates of

convergence that are independent of ω .

APPENDIX A: PROOF OF THEOREM 3.1

To prove Theorem 3.1, we first relate the objective functions of (3.1) and (3.2), and the corresponding $R_\omega(\phi)$ and $L_\omega(\phi)$, to the support functions $h_{\mathcal{F}_{\mu\nu}^A}(\cdot)$ and $h_{\mathcal{F}_{\mu\nu}^\omega}(\cdot)$. The result then follows directly from Theorem 4.1.

LEMMA A.1. *Assume the notation of Theorem 3.1, and let $R_A(\phi)$ be the objective function of (3.1). Then, given $\phi \equiv (\mu, \nu, \theta) \in \Phi$,*

$$R_A(\phi) - R_\omega(\phi) = 2[h_{\mathcal{F}_{\mu\nu}^\omega}(\theta) - h_{\mathcal{F}_{\mu\nu}^A}(\theta)] + \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n A_{ij}^2 - \int_{[0,1]^2} \omega(x, y)^2 dx dy.$$

Let $L_A(\phi)$ be the objective function of (3.2). Define $B_\theta \in \mathbb{R}^+$, $\Gamma_\theta \in [-1, 1]^{K \times K}$ to satisfy $B_\theta \Gamma_\theta(a, b) = \log(\theta_{ab}/(1 - \theta_{ab}))$ for $a, b = 1, \dots, K$. Then

$$L_A(\phi) - L_\omega(\phi) = B_\theta [h_{\mathcal{F}_{\mu\nu}^A}(\Gamma_\theta) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma_\theta)].$$

To complete the proof of the first part of Theorem 3.1, let $\hat{\phi} \equiv (\hat{\mu}, \hat{\nu}, \hat{\theta}) = \operatorname{argmin}_{\phi \in \Phi} R_A(\phi)$. For any $\phi \equiv (\mu, \nu, \theta)$ in Φ ,

$$\begin{aligned} R_\omega(\hat{\phi}) - R_\omega(\phi) &= R_\omega(\hat{\phi}) - R_A(\hat{\phi}) + R_A(\hat{\phi}) - R_A(\phi) + R_A(\phi) - R_\omega(\phi) \\ &\leq R_\omega(\hat{\phi}) - R_A(\hat{\phi}) + R_A(\phi) - R_\omega(\phi) \\ &\leq 2|h_{\mathcal{F}_{\mu\nu}^A}(\hat{\theta}) - h_{\mathcal{F}_{\mu\nu}^\omega}(\hat{\theta})| + 2|h_{\mathcal{F}_{\mu\nu}^\omega}(\theta) - h_{\mathcal{F}_{\mu\nu}^A}(\theta)|, \end{aligned}$$

where the first inequality holds because $R_A(\hat{\phi}) - R_A(\phi) \leq 0$, and the second holds by the triangle inequality and Lemma A.1. Applying Theorem 4.1 and choosing ϕ to satisfy $R_\omega(\phi) \leq \inf_{\phi' \in \Phi} R_\omega(\phi') + n^{-1/4}$ then yields the claimed result that $R_\omega(\hat{\phi}) - \inf_{\phi \in \Phi} R_\omega(\phi) = \mathcal{O}_P(n^{-1/4})$.

Now set $\bar{\phi} = \operatorname{argmax}_{\phi \in \Phi} L_\omega(\phi)$ and $\hat{\phi} = \operatorname{argmax}_{\phi \in \Phi} L_A(\phi)$; the result $[L_\omega(\bar{\phi}) - L_\omega(\hat{\phi})]/[B(\bar{\theta}) + B(\hat{\theta})] = \mathcal{O}_P(n^{-1/4})$ follows similarly from

$$\begin{aligned} 0 \leq L_\omega(\bar{\phi}) - L_\omega(\hat{\phi}) &= L_\omega(\bar{\phi}) - L_A(\bar{\phi}) + L_A(\bar{\phi}) - L_A(\hat{\phi}) + L_A(\hat{\phi}) - L_\omega(\hat{\phi}) \\ &\leq L_\omega(\bar{\phi}) - L_A(\bar{\phi}) + L_A(\hat{\phi}) - L_\omega(\hat{\phi}) \\ &\leq B(\bar{\theta})|h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma_{\bar{\theta}}) - h_{\mathcal{F}_{\mu\nu}^A}(\Gamma_{\bar{\theta}})| + B(\hat{\theta})|h_{\mathcal{F}_{\mu\nu}^A}(\Gamma_{\hat{\theta}}) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma_{\hat{\theta}})|. \end{aligned}$$

PROOF OF LEMMA A.1. We show the results of the lemma directly:

$$\begin{aligned}
R_A(\phi) &= \min_{(S,T) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n |\theta_{S(i)T(j)} - A_{ij}|^2 \\
&= \min_{F \in \mathcal{F}_{\mu\nu}^A} \left\{ \sum_{a=1}^K \sum_{b=1}^K -2F_{ab}\theta_{ab} + \mu_a\nu_b\theta_{ab}^2 \right\} + \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n A_{ij}^2 \\
&= \left\{ -2h_{\mathcal{F}_{\mu\nu}^A}(\theta) + \sum_{a=1}^K \sum_{b=1}^K \mu_a\nu_b\theta_{ab}^2 \right\} + \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n A_{ij}^2,
\end{aligned}$$

where the second line follows from the definition of $\mathcal{F}_{\mu\nu}^A$, and the last line from that of $h_{\mathcal{F}_{\mu\nu}^A}$. To complete the first result of the lemma, observe that by letting σ and τ satisfy $\sigma(x) = F_\mu^{-1}(\pi_1(x))$ and $\tau(y) = F_\nu^{-1}(\pi_2(y))$,

$$\begin{aligned}
R_\omega(\phi) &= \inf_{\pi_1, \pi_2 \in \mathcal{P}} \int_{[0,1]^2} |\omega(\pi_1(x), \pi_2(y)) - \omega_\phi(x, y)|^2 dx dy \\
&= \inf_{(\sigma, \tau) \in \mathcal{Q}_\mu \times \mathcal{Q}_\nu} \sum_{a=1}^K \sum_{b=1}^K \int_{\sigma^{-1}(a) \times \tau^{-1}(b)} |\omega(x, y) - \theta_{ab}|^2 dx dy \\
&= \inf_{F \in \mathcal{F}_{\mu\nu}^\omega} \left\{ \sum_{a=1}^K \sum_{b=1}^K -2F_{ab}\theta_{ab} + \mu_a\nu_b\theta_{ab}^2 \right\} + \int_{[0,1]^2} \omega(x, y)^2 dx dy \\
&= \left\{ -2h_{\mathcal{F}_{\mu\nu}^\omega}(\theta) + \sum_{a=1}^K \sum_{b=1}^K \mu_a\nu_b\theta_{ab}^2 \right\} + \int_{[0,1]^2} \omega(x, y)^2 dx dy.
\end{aligned}$$

Following similar steps, we show the second result as follows:

$$\begin{aligned}
L_A(\phi) &= \max_{(S,T) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n \{A_{ij} \log(\theta_{S(i)T(j)}) + (1 - A_{ij}) \log(1 - \theta_{S(i)T(j)})\} \\
&= \max_{F \in \mathcal{F}_{\mu\nu}^A} \sum_{a=1}^K \sum_{b=1}^K \left\{ F_{ab} \log\left(\frac{\theta_{ab}}{1 - \theta_{ab}}\right) + \mu_a\nu_b \log(1 - \theta_{ab}) \right\} \\
&= B_\theta h_{\mathcal{F}_{\mu\nu}^A}(\Gamma_\theta) + \sum_{a=1}^K \sum_{b=1}^K \mu_a\nu_b \log(1 - \theta_{ab}),
\end{aligned}$$

since $\max_{F \in \mathcal{F}_{\mu\nu}^A} \sum_{a,b} F_{ab} B_\theta \Gamma_\theta(a, b) = B_\theta h_{\mathcal{F}_{\mu\nu}^A}(\Gamma_\theta)$, and similarly

$$\begin{aligned}
L_\omega(\phi) &= \sup_{\pi_1, \pi_2 \in \mathcal{P}} \int_{[0,1]^2} \{ \omega(\pi_1(x), \pi_2(y)) \log \omega_\phi(x, y) \\
&\quad + [1 - \omega(\pi_1(x), \pi_2(y))] \log(1 - \omega_\phi(x, y)) \} dx dy, \\
&= \sup_{(\sigma, \tau) \in \mathcal{Q}_\mu \times \mathcal{Q}_\nu} \sum_{a=1}^K \sum_{b=1}^K \int_{\sigma^{-1}(a) \times \tau^{-1}(b)} \{ \omega(x, y) \log \theta_{ab} \\
&\quad + (1 - \omega(x, y)) \log(1 - \theta_{ab}) \} dx dy \\
&= \sup_{F \in \mathcal{F}_{\mu\nu}^\omega} \sum_{a=1}^K \sum_{b=1}^K \left\{ F_{ab} \log \left(\frac{\theta_{ab}}{1 - \theta_{ab}} \right) + \mu_a \nu_b \log(1 - \theta_{ab}) \right\} \\
&= B_\theta h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma_\theta) + \sum_{a=1}^K \sum_{b=1}^K \mu_a \nu_b \log(1 - \theta_{ab}).
\end{aligned}$$

□

APPENDIX B: AUXILIARY PROOFS FOR THEOREM 4.1

Below we provide proofs of all supporting lemmas for Theorem 4.1, and state and prove the covering argument used to establish the theorem.

1. First, in Sections B.1–B.3 below, we prove auxiliary Lemmas 5.2, 5.4, and 5.5 as stated in Section 5.
2. Then, in Section B.4, we generalize the definitions of \mathcal{Q}_U and \mathcal{Q}_V , given in Section 5.2 for $K = 2$, to arbitrary K ; this induces generalizations of the quantities δ_U, δ_V , and δ_{UV} in the natural way.
3. Then, in Sections B.5 and B.6, we prove Lemmas 5.6 and 5.7, which depend on $(\mathcal{Q}_U, \mathcal{Q}_V, \delta_U, \delta_V, \delta_{UV})$ as defined for arbitrary K .
4. Finally, in Section B.7, we extend the pointwise convergence result of Proposition 5.1 by way of a covering argument for all $\Gamma \in [-1, 1]^{K \times K}$.

B.1. Proof of Lemma 5.2. For fixed Γ , let $(\sigma^*, \tau^*) \in \mathcal{Q}_\mu \times \mathcal{Q}_\nu$ satisfy

$$(B.1) \quad \langle \Gamma, \omega / \sigma^* \tau^* \rangle > h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma) - \frac{1}{n^{1/4}},$$

so that $\omega / \sigma^* \tau^*$ is within $n^{-1/4}$ of the supporting hyperplane. Define

$$S^*(i) = \sigma^*(\xi_i), \quad T^*(j) = \tau^*(\zeta_j); \quad i = 1, \dots, m, \quad j = 1, \dots, n.$$

By the arguments of Lemma 5.4 as proved in Section B.2 below, applying McDiarmid's inequality with the Lipschitz conditions of Lemma 5.1 yields

$$(B.2) \quad \mathbb{P}(|\langle \Gamma, A / S^* T^* \rangle - \langle \Gamma, \omega / \sigma^* \tau^* \rangle| \geq 2\epsilon) \leq 2e^{-2mn\epsilon^2/(m+n)} + 2e^{-2mn\epsilon^2}.$$

While (S^*, T^*) may not be in $\mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n$, a Chernoff bound implies that

$$\mathbb{P}\left(\left|\frac{S^{*-1}(a)}{m} - \mu_a\right| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}, \quad a = 1, \dots, K.$$

The analogous bound also holds for $|T^{*-1}(b)/n - \nu_b|$. Applying these results in conjunction with a union bound yields

$$\mathbb{P}\left(\max_{1 \leq a, b \leq K} \left\{\left|\frac{S^{*-1}(a)}{m} - \mu_a\right| + \left|\frac{T^{*-1}(b)}{n} - \nu_b\right|\right\} \geq 2\epsilon\right) \leq K(2e^{-2m\epsilon^2} + 2e^{-2n\epsilon^2}).$$

Therefore, with probability at least $1 - K(2e^{-2m\epsilon^2} + 2e^{-2n\epsilon^2})$, there exists a pair $(\mathring{S}, \mathring{T}) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n$ such that

$$\frac{1}{m}d_{\text{Ham}}(S^*, \mathring{S}) + \frac{1}{n}d_{\text{Ham}}(T^*, \mathring{T}) \leq 2K\epsilon,$$

which by the first condition of Lemma 5.1 implies that

$$(B.3) \quad |\langle \Gamma, A/\mathring{S}\mathring{T} \rangle - \langle \Gamma, A/S^*T^* \rangle| \leq 4K\epsilon.$$

Recalling that $h_{\mathcal{F}_{\mu\nu}^A} = \max_{(S,T) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \langle \Gamma, A/ST \rangle$, we have that

$$h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) \geq \langle \Gamma, A/\mathring{S}\mathring{T} \rangle,$$

following which (B.3), (B.2), and (B.1) in turn imply that with probability at least $1 - 2e^{-2mn\epsilon^2/(m+n)} - 2e^{-2mn\epsilon^2} - K(2e^{-2m\epsilon^2} + 2e^{-2n\epsilon^2})$, we have $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma)$

$$\begin{aligned} &\geq \langle \Gamma, A/S^*T^* \rangle - 4K\epsilon \\ &\geq \langle \Gamma, \omega/\sigma^*\tau^* \rangle - (4K + 2)\epsilon \\ &\geq h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma) - n^{-1/4} - (4K + 2)\epsilon. \end{aligned}$$

Now letting $m = \rho n$ as in the statement of the lemma, and setting $\epsilon = n^{-1/4}$, we see that with probability at least $1 - 2e^{-\sqrt{n}[2\rho/(\rho+1)]} [1 + o(1)]$,

$$h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) \geq h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma) - \frac{4K + 3}{n^{1/4}},$$

providing a lower bound on $h_{\mathcal{F}_{\mu\nu}^A}(\Gamma)$ that converges to $h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)$.

B.2. Proof of Lemma 5.4. Recalling the definitions of $h_{\mathcal{F}_{\mu\nu}^A}$ and $h_{\mathcal{F}_{\mu\nu}^W}$,

$$\begin{aligned}
 & \mathbb{P} \left(|h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)| \geq \epsilon \right) \\
 &= \mathbb{P} \left(\left| \max_{(S,T) \in \mathcal{Q}_{\mu}^m \times \mathcal{Q}_{\nu}^n} \langle \Gamma, A/ST \rangle - \max_{(S,T) \in \mathcal{Q}_{\mu}^m \times \mathcal{Q}_{\nu}^n} \langle \Gamma, W/ST \rangle \right| \geq \epsilon \right) \\
 &\leq \mathbb{P} \left(\max_{(S,T) \in \mathcal{Q}_{\mu}^m \times \mathcal{Q}_{\nu}^n} |\langle \Gamma, A/ST \rangle - \langle \Gamma, W/ST \rangle| \geq \epsilon \right) \\
 \text{(B.4)} \quad &\leq \sum_{(S,T) \in \mathcal{Q}_{\mu}^m \times \mathcal{Q}_{\nu}^n} \mathbb{P} (|\langle \Gamma, A/ST \rangle - \langle \Gamma, W/ST \rangle| \geq \epsilon),
 \end{aligned}$$

where (B.4) follows from a union bound.

Now consider $\langle \Gamma, A/ST \rangle$ as a function of the mn independent random variables $\{A_{ij}\}$, and observe that $\mathbb{E}(\langle \Gamma, A/ST \rangle) = \langle \Gamma, W/ST \rangle$ for each (S, T) , since $W_{ij} = \omega(\xi_i, \zeta_j) = \mathbb{E}(A_{ij})$. Also recall the final Lipschitz condition of Lemma 5.1, which states that $|\langle \Gamma, A/ST \rangle - \langle \Gamma, A'/ST \rangle| \leq 1/(mn)$ if A and A' differ by a single entry. Thus we may apply McDiarmid's inequality to $\langle \Gamma, A/ST \rangle$ in the summand of (B.4), and since $|\mathcal{Q}_{\mu}^m| \leq K^m$ and $|\mathcal{Q}_{\nu}^n| \leq K^n$,

$$\mathbb{P} \left(|h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)| \geq \epsilon \right) \leq K^{m+n} \cdot 2e^{-2mn\epsilon^2}.$$

Now consider $h_{\mathcal{F}_{\mu\nu}^W}(\Gamma) = \max_{(S,T) \in \mathcal{Q}_{\mu}^m \times \mathcal{Q}_{\nu}^n} \langle \Gamma, W/ST \rangle$ as a function of the $m+n$ independent random variables ξ_1, \dots, ξ_m and ζ_1, \dots, ζ_n . Changing a single component of ξ or ζ affects a single row or column of W , respectively, and thus alters $\langle \Gamma, W/ST \rangle$ and hence $h_{\mathcal{F}_{\mu\nu}^W}$ by at most $1/m$ or $1/n$. It therefore follows from McDiarmid's inequality that

$$\mathbb{P} \left(|h_{\mathcal{F}_{\mu\nu}^W}(\Gamma) - \mathbb{E} h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)| \geq \epsilon \right) \leq 2e^{-2mn\epsilon^2/(m+n)}.$$

Combining these inequalities via a union bound yields the statement of the lemma, since by the triangle inequality we must have $|h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)| \geq \epsilon$ or $|h_{\mathcal{F}_{\mu\nu}^W}(\Gamma) - \mathbb{E} h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)| \geq \epsilon$ in order that $|h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - \mathbb{E} h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)| \geq 2\epsilon$.

B.3. Proof of Lemma 5.5. Recall from the statement of the lemma that \mathcal{I} and \mathcal{J} denote sets of deterministic size whose elements are sampled without replacement from $1, \dots, m$ and $1, \dots, n$, respectively. We adopt the notation that $\mathbb{E}_{\mathcal{I}}$ denotes an expectation taken over \mathcal{I} , with all other random variables held constant, and define $\mathbb{E}_{\mathcal{J}}$ and $\mathbb{E}_{\mathcal{I}\mathcal{J}}$ in the same manner.

To prove the lemma, it suffices to show that for all W, T, S ,

$$(B.5) \quad \mathbb{E}_{\mathcal{J}} \left(\langle \Gamma, W / \hat{S}^T T \rangle \right) \geq \langle \Gamma, W / S^T T \rangle - K \sqrt{2\pi / |\mathcal{J}|}$$

$$(B.6) \quad \mathbb{E}_{\mathcal{I}} \left(\langle \Gamma, W / S \hat{T}^S \rangle \right) \geq \langle \Gamma, W / S T^S \rangle - K \sqrt{2\pi / |\mathcal{I}|},$$

where \hat{S}^T and \hat{T}^S are respectively defined in (5.2) and (5.3), and

$$S^T = \operatorname{argmax}_{S \in \mathcal{Q}_{\mu}^m} \langle \Gamma, W / S T \rangle, \quad T^S = \operatorname{argmax}_{T \in \mathcal{Q}_{\nu}^n} \langle \Gamma, W / S T \rangle.$$

This is because (B.5) and (B.6) imply that for all $(U, V) \in \mathcal{Q}_{\mu}^m \times \mathcal{Q}_{\nu}^n$,

$$\begin{aligned} \langle \Gamma, W / UV \rangle &\leq \langle \Gamma, W / U T^U \rangle \\ &\leq \mathbb{E}_{\mathcal{I}} \left(\langle \Gamma, W / U \hat{T}^U \rangle \right) + K \sqrt{2\pi / |\mathcal{I}|} \\ &\leq \mathbb{E}_{\mathcal{I}} \left(\langle \Gamma, W / S^{\hat{T}^U} \hat{T}^U \rangle \right) + K \sqrt{2\pi / |\mathcal{I}|} \\ &\leq \mathbb{E}_{\mathcal{I}} \mathbb{E}_{\mathcal{J}} \left(\langle \Gamma, W / \hat{S}^{\hat{T}^U} \hat{T}^U \rangle \right) + K \sqrt{2\pi / |\mathcal{I}|} + K \sqrt{2\pi / |\mathcal{J}|} \\ &\leq \mathbb{E}_{\mathcal{I}\mathcal{J}} \left(\max_{(Q, R) \in \mathcal{Q}_{\mu}^m \times \mathcal{Q}_{\nu}^n} \langle \Gamma, W / \hat{S}^R \hat{T}^Q \rangle \right) + K \sqrt{2\pi} \left(|\mathcal{I}|^{-\frac{1}{2}} + |\mathcal{J}|^{-\frac{1}{2}} \right). \end{aligned}$$

Recalling the definition of $h_{\mathcal{F}_{\mu\nu}^W}(\Gamma)$, and noting that the right-hand side above is deterministic for fixed W , with no dependence on U or V , we may write

$$\begin{aligned} h_{\mathcal{F}_{\mu\nu}^W}(\Gamma) &= \max_{(U, V) \in \mathcal{Q}_{\mu}^m \times \mathcal{Q}_{\nu}^n} \langle \Gamma, W / UV \rangle \\ &\leq \mathbb{E}_{\mathcal{I}\mathcal{J}} \left(\max_{(Q, R) \in \mathcal{Q}_{\mu}^m \times \mathcal{Q}_{\nu}^n} \langle \Gamma, W / \hat{S}^R \hat{T}^Q \rangle \right) + K \sqrt{2\pi} \left(|\mathcal{I}|^{-\frac{1}{2}} + |\mathcal{J}|^{-\frac{1}{2}} \right). \end{aligned}$$

Taking expectations on both sides over W gives the statement of the lemma.

We now establish (B.5), noting that (B.6) will follow by parallel arguments. For fixed W and T , define for any $a = 1, \dots, K$ the difference

$$\Delta_i^a = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} W_{ij} \Gamma_{aT(j)} - \frac{1}{n} \sum_{j=1}^n W_{ij} \Gamma_{aT(j)}.$$

It follows that $\mathbb{E}_{\mathcal{J}}(\Delta_i^a) = 0$, and by a Chernoff bound,

$$\mathbb{P}(|\Delta_i^a| \geq t) \leq 2e^{-2t^2|\mathcal{J}|}.$$

As $|\Delta_i^a|$ is nonnegative, the identity $\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X \geq t) dt$ for X taking only nonnegative values can be used to bound its expectation according to

$$\mathbb{E}_{\mathcal{J}}(|\Delta_i^a|) \leq \sqrt{\pi / (2|\mathcal{J}|)},$$

which implies

$$(B.7) \quad \mathbb{E}_{\mathcal{J}} \left(\max_{1 \leq a \leq K} |\Delta_i^a| \right) \leq K \sqrt{\pi/(2|\mathcal{J}|)}.$$

For fixed W and \mathcal{J} , define the function

$$f_W(S, T) = \frac{1}{m|\mathcal{J}|} \sum_{i=1}^m \sum_{j \in \mathcal{J}} W_{ij} \Gamma_{S(i)T(j)},$$

and for fixed W and T , let

$$(B.8) \quad \Delta = \max_{S \in \mathcal{Q}_{\mu}^m} |f_W(S, T) - \langle \Gamma, W/ST \rangle|.$$

From the definition of Δ it follows that

$$\begin{aligned} \Delta &= \max_{S \in \mathcal{Q}_{\mu}^m} \left\{ \frac{1}{m} \left| \sum_{i=1}^m \left(\frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} W_{ij} \Gamma_{S(i)T(j)} - \frac{1}{n} \sum_{j=1}^n W_{ij} \Gamma_{S(i)T(j)} \right) \right| \right\} \\ &\leq \frac{1}{m} \left| \sum_{i=1}^m \max_{1 \leq a \leq K} \left\{ \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} W_{ij} \Gamma_{aT(j)} - \frac{1}{n} \sum_{j=1}^n W_{ij} \Gamma_{aT(j)} \right\} \right| \\ &= \frac{1}{m} \left| \sum_{i=1}^m \max_{1 \leq a \leq K} \{\Delta_i^a\} \right| \leq \frac{1}{m} \sum_{i=1}^m \max_{1 \leq a \leq K} |\Delta_i^a|. \end{aligned}$$

Taking expectations of both sides over \mathcal{J} and substituting (B.7) yields

$$(B.9) \quad \mathbb{E}_{\mathcal{J}}(\Delta) \leq \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{\mathcal{J}} \left(\max_{1 \leq a \leq K} |\Delta_i^a| \right) \leq K \sqrt{\pi/(2|\mathcal{J}|)}.$$

Finally, to show (B.5), observe that since \hat{S}^T from (5.2) maximizes $f_W(\cdot, T)$, and S^T as defined above maximizes $\langle \Gamma, W/\cdot T \rangle$, we have from (B.8) that

$$\begin{aligned} 0 &\leq \langle \Gamma, W/S^T T \rangle - \langle \Gamma, W/\hat{S}^T T \rangle \\ &\leq \langle \Gamma, W/S^T T \rangle - f_W(S^T, T) + f_W(\hat{S}^T, T) - \langle \Gamma, W/\hat{S}^T T \rangle \leq 2\Delta, \end{aligned}$$

and so $\langle \Gamma, W/\hat{S}^T T \rangle \geq \langle \Gamma, W/S^T T \rangle - 2\Delta$. Taking expectations of both sides of this expression over \mathcal{J} , and then substituting (B.9), yields the inequality

$$\mathbb{E}_{\mathcal{J}} \left(\langle \Gamma, W/\hat{S}^T T \rangle \right) \geq \langle \Gamma, W/S^T T \rangle - 2K \sqrt{\pi/(2|\mathcal{J}|)},$$

which is the statement of (B.5). That of (B.6) follows by parallel arguments.

B.4. Definition of \mathcal{Q}_U and \mathcal{Q}_V for arbitrary K . In order to redefine \mathcal{Q}_U and \mathcal{Q}_V to accommodate arbitrary K , we first redefine the mappings U and V . Given $\zeta_{\mathcal{J}} = \{\zeta_j : j \in \mathcal{J}\}$ and an assignment function $R : \{1, \dots, n\} \rightarrow \{1, \dots, K\}$, define the mapping $U : [0, 1] \rightarrow \mathbb{R}^K$ by

$$U_a(x) = \sum_{j \in \mathcal{J}} \omega(x, \zeta_j) \Gamma_{aR(j)}; \quad x \in [0, 1], \quad a = 1, \dots, K.$$

Analogously, given $\xi_{\mathcal{I}}$ and Q , define $V : [0, 1] \rightarrow \mathbb{R}^K$ by

$$V_a(y) = \sum_{i \in \mathcal{I}} \omega(\xi_i, y) \Gamma_{Q(i)a}, \quad y \in [0, 1], \quad a = 1, \dots, K.$$

Given $a, b \in \{1, \dots, K\}$ and the mapping U , define the relation $\succeq^{U,a,b}$ by

$$x_1 \succeq^{U,a,b} x_2 \equiv \begin{cases} U_a(x_1) - U_b(x_2) > U_a(x_2) - U_b(x_1), & \text{or} \\ U_a(x_1) - U_b(x_2) = U_a(x_2) - U_b(x_1), & \text{if } (a-b)(x_1-x_2) \geq 0. \end{cases}$$

Informally, $x_1 \succeq^{U,a,b} x_2$ implies that, given the choice of assigning either x_1 or x_2 to group a , with the other relegated to group b , x_1 is at least as attractive as x_2 . The latter tie-breaker condition results in a symmetric definition: if $x_1 \succeq^{U,a,b} x_2$, then $x_2 \succeq^{U,b,a} x_1$. We define $\succ^{U,a,b}$ analogously to $\succeq^{U,a,b}$, except that the inequality $(a-b)(x_1-x_2) > 0$ is strict.

Let \mathcal{S} denote the set of symmetric matrices in $[0, 1]^{K \times K}$. Given $t \in \mathcal{S}$ and the mapping U , we define the function $\sigma_t : [0, 1] \rightarrow \{1, \dots, K\}$ as the mapping which satisfies the following

$$\sigma_t^{-1}(a) = \{x : x \succeq^{U,a,b} t_{ab} \forall b > a, x \succ^{U,a,b} t_{ab} \forall b < a\}, \quad a = 1, \dots, K,$$

with the convention that σ_t is undefined whenever the above rule does not map all of $[0, 1]$ to $\{1, \dots, K\}$.

We define the function class \mathcal{Q}_U as follows:

$$\mathcal{Q}_U = \{\sigma_t : t \in \mathcal{S} \text{ and } \sigma_t \text{ is defined}\}.$$

Given $t \in \mathcal{S}$ and the mapping V as defined above, we define $\succ^{V,a,b}$, τ_t , and \mathcal{Q}_V analogously. We then have the following.

LEMMA B.1. *Given U induced by $\zeta_{\mathcal{J}}$ and R , and given W induced by ξ and ζ , define \hat{S}^R by (5.2). Then there exists $\hat{\sigma} \in \mathcal{Q}_U$ such that*

$$\hat{S}^R(i) = \hat{\sigma}(\xi_i), \quad i = 1, \dots, m.$$

Likewise, given V induced by $\xi_{\mathcal{I}}$ and Q , and given W induced by ξ and ζ , define \hat{T}^Q by (5.3). Then there exists $\hat{\tau} \in \mathcal{Q}_V$ such that

$$\hat{T}^Q(j) = \hat{\tau}(\zeta_j), \quad j = 1, \dots, n.$$

PROOF OF LEMMA B.1. Let \hat{S}^R be chosen lexicographically from the set of all maximizers of (5.2), where S lexicographically precedes S' if and only if $S(i_1), \dots, S(i_m)$ lexicographically precedes $S'(i_1), \dots, S'(i_m)$, where i_1, \dots, i_m are in order of increasing $\xi_{i_1}, \dots, \xi_{i_m}$.

Since \hat{S}^R maximizes (5.2), it holds for all $i, j = 1, \dots, m$ that

$$U_{\hat{S}^R(i)}(\xi_i) + U_{\hat{S}^R(j)}(\xi_j) \geq U_{\hat{S}^R(i)}(\xi_j) + U_{\hat{S}^R(j)}(\xi_i);$$

otherwise switching labels for i and j would increase the value of the objective function. As \hat{S}^R is chosen lexicographically, for any i, j such that

$$U_{\hat{S}^R(i)}(\xi_i) + U_{\hat{S}^R(j)}(\xi_j) = U_{\hat{S}^R(i)}(\xi_j) + U_{\hat{S}^R(j)}(\xi_i),$$

it holds that $(\hat{S}^R(i) - \hat{S}^R(j))(\xi_i - \xi_j) \geq 0$, with equality if and only if $\xi_i = \xi_j$. Otherwise, switching labels would improve the lexicographic ordering.

Since $\xi_i \neq \xi_j$ for $i \neq j$ except on a set of measure zero, it follows that

$$(\hat{S}^R)^{-1}(a) \succ^{U,a,b} (\hat{S}^R)^{-1}(b), \quad a, b = 1, \dots, K, \quad a \neq b,$$

where we have let $(\hat{S}^R)^{-1}(a)$ denote $\{\xi_i : \hat{S}^R(\xi_i) = a\}$. As a result, for each a and b we may choose $t_{ab} = t_{ba} \in [0, 1]$ such that $(\hat{S}^R)^{-1}(a) \succ^{U,a,b} t_{ab}$ and $(\hat{S}^R)^{-1}(b) \succ^{U,b,a} t_{ba}$, implying that $\hat{S}^R(i) = \hat{\sigma}(\xi_i)$ for some $\hat{\sigma} \in \mathcal{Q}_U$. As parallel arguments hold for \hat{T}^Q , the statement of the lemma follows. \square

B.5. Proof of Lemma 5.6. Recall the definition of δ_{UV} from Section 5.2, which we can now interpret for arbitrary K according to the definitions of \mathcal{Q}_U and \mathcal{Q}_V in Section B.4 above. We use a symmetrization argument of Hoeffding [11, 17] to bound $\mathbb{E}(\max_{(Q,R) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n} \delta_{UV})$. Let $\mathcal{M}_{\mathcal{I}}$ denote the set of permutations of $1, \dots, m$ which map $1, \dots, m - |\mathcal{I}|$ to $i \notin \mathcal{I}$, and let $\mathcal{M}_{\mathcal{J}}$ be defined analogously for permutations on $1, \dots, n$. Let $\mathcal{M} = \mathcal{M}_{\mathcal{I}} \times \mathcal{M}_{\mathcal{J}}$ and let $Z = |\mathcal{M}|$. Let ξ', ζ' be identically distributed as ξ and ζ , and independent of U and V . Let $\xi_{\mathcal{I}}$ and $\zeta_{\mathcal{J}}$ be defined as in Section B.4. To abbreviate the notation, let $g_{\sigma\tau}(x, y) = \omega(x, y) \Gamma_{\sigma(x)\tau(y)}$, and let $\mathcal{Q} = \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n \times \mathcal{Q}_U \times \mathcal{Q}_V$. It holds for $(Q, R) \in \mathcal{Q}_\mu^m \times \mathcal{Q}_\nu^n$ that

$$\mathbb{E} \left(\max_{Q,R} \delta_{UV} \right) = \mathbb{E} \left(\sup_{(Q,R,\sigma,\tau) \in \mathcal{Q}} |G_{\sigma,\tau}(\xi, \zeta) - \mathbb{E}(G_{\sigma\tau}(\xi', \zeta') | U, V)| | \xi_{\mathcal{I}}, \zeta_{\mathcal{J}} \right),$$

which by convexity can be upper-bounded by

$$\begin{aligned}
& \mathbb{E} \left(\sup_{(Q,R,\sigma,\tau) \in \mathcal{Q}} |G_{\sigma,\tau}(\xi, \zeta) - G_{\sigma,\tau}(\xi', \zeta')| \mid \xi_{\mathcal{I}}, \zeta_{\mathcal{J}} \right) \\
&= \mathbb{E} \left(\sup_{(Q,R,\sigma,\tau) \in \mathcal{Q}} \left| \frac{1}{mn} \sum_{i \notin \mathcal{I}} \sum_{j \notin \mathcal{J}} g_{\sigma,\tau}(\xi_i, \zeta_j) - g_{\sigma,\tau}(\xi'_i, \zeta'_j) \right| \mid \xi_{\mathcal{I}}, \zeta_{\mathcal{J}} \right) \\
&= \mathbb{E} \left(\sup_{(Q,R,\sigma,\tau) \in \mathcal{Q}} \left| \frac{|\bar{\mathcal{I}}||\bar{\mathcal{J}}|}{Zmn} \sum_{\pi, \eta \in \mathcal{M}} \frac{1}{\ell} \sum_{i=1}^{\ell} g_{\sigma,\tau}(\xi_{\pi(i)}, \zeta_{\eta(j)}) - g_{\sigma,\tau}(\xi'_{\pi(i)}, \zeta'_{\eta(j)}) \right| \mid \xi_{\mathcal{I}}, \zeta_{\mathcal{J}} \right),
\end{aligned}$$

since the permutations π and η weight each (i, j) term equally for $i \notin \mathcal{I}$ and $j \notin \mathcal{J}$; by convexity again, and then linearity of expectation, we have

$$\begin{aligned}
& \leq \mathbb{E} \left(\frac{|\bar{\mathcal{I}}||\bar{\mathcal{J}}|}{Zmn} \sum_{\pi, \eta \in \mathcal{M}} \sup_{(Q,R,\sigma,\tau) \in \mathcal{Q}} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} g_{\sigma,\tau}(\xi_{\pi(i)}, \zeta_{\eta(j)}) - g_{\sigma,\tau}(\xi'_{\pi(i)}, \zeta'_{\eta(j)}) \right| \mid \xi_{\mathcal{I}}, \zeta_{\mathcal{J}} \right) \\
&= \frac{|\bar{\mathcal{I}}||\bar{\mathcal{J}}|}{mn} \mathbb{E} \left(\sup_{(Q,R,\sigma,\tau) \in \mathcal{Q}} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} g_{\sigma,\tau}(\xi_i, \zeta_i) - g_{\sigma,\tau}(\xi'_i, \zeta'_i) \right| \mid \xi_{\mathcal{I}}, \zeta_{\mathcal{J}} \right).
\end{aligned}$$

We may now introduce independent and identically distributed Rademacher variables r_1, \dots, r_{ℓ} , and use standard Rademacher symmetrization arguments (see, e.g., [8]) to show that the final quantity above is equal to

$$\begin{aligned}
& \frac{|\bar{\mathcal{I}}||\bar{\mathcal{J}}|}{mn} \mathbb{E} \left(\sup_{(Q,R,\sigma,\tau) \in \mathcal{Q}} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} r_i (g_{\sigma,\tau}(\xi_i, \zeta_i) - g_{\sigma,\tau}(\xi'_i, \zeta'_i)) \right| \mid \xi_{\mathcal{I}}, \zeta_{\mathcal{J}} \right) \\
& \leq \frac{|\bar{\mathcal{I}}||\bar{\mathcal{J}}|}{mn} \mathbb{E} \left(\sup_{(Q,R,\sigma,\tau) \in \mathcal{Q}} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} r_i g_{\sigma,\tau}(\xi_i, \zeta_i) \right| + \left| \frac{1}{\ell} \sum_{i=1}^{\ell} r_i g_{\sigma,\tau}(\xi'_i, \zeta'_i) \right| \mid \xi_{\mathcal{I}}, \zeta_{\mathcal{J}} \right) \\
& \leq 2 \frac{|\bar{\mathcal{I}}||\bar{\mathcal{J}}|}{mn} \mathbb{E} \left(\sup_{(Q,R,\sigma,\tau) \in \mathcal{Q}} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} r_i g_{\sigma,\tau}(\xi_i, \zeta_i) \right| \mid \xi_{\mathcal{I}}, \zeta_{\mathcal{J}} \right).
\end{aligned}$$

To bound this expectation, note that for fixed $\mathcal{I}, \mathcal{J}, Q, R$ (inducing a fixed U and V), and fixed $(\sigma, \tau) \in \mathcal{Q}_U \times \mathcal{Q}_V$, a Hoeffding inequality gives

$$\text{(B.10)} \quad \mathbb{P} \left(\left| \frac{1}{\ell} \sum_{i=1}^{\ell} r_i g_{\sigma,\tau}(\xi_i, \zeta_j) \right| \geq \epsilon \mid \xi_{\mathcal{I}}, \zeta_{\mathcal{J}} \right) \leq 2e^{-2\ell\epsilon^2}.$$

We may now apply (B.10) in conjunction with a union bound over all $(Q, R, \sigma, \tau) \in \mathcal{Q}$ as follows. For fixed Q, R, a, b , the set $\{i : \xi_i \succeq^{U,a,b} t_{ab}\}$ can

be chosen at most $\ell + 1$ ways by varying t_{ab} . As a result, the set ξ_1, \dots, ξ_ℓ can be partitioned at most $(\ell + 1)^{\binom{K}{2}}$ ways by varying $\sigma \in \mathcal{Q}_U$. Analogously, the set $\zeta_1, \dots, \zeta_\ell$ can be partitioned the same number of ways by varying $\tau \in \mathcal{Q}_V$. For fixed \mathcal{I}, \mathcal{J} , the functions U and V can be chosen $K^{|\mathcal{I}|+|\mathcal{J}|}$ different ways by varying Q and R . Hence, a union bound gives

$$\mathbb{P} \left(\sup_{(Q,R,\sigma,\tau) \in \mathcal{Q}} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} r_i g_{\sigma\tau}(\xi_i \zeta_i) \right| \geq \epsilon \mid \xi_{\mathcal{I}}, \zeta_{\mathcal{J}} \right) \leq K^{|\mathcal{I}|+|\mathcal{J}|} (\ell+1)^{2\binom{K}{2}} \cdot 2e^{-2\ell\epsilon^2}.$$

Since this expression is of the form $\mathbb{P}(X \geq t) \leq f(t)$ for X nonnegative, we may apply the inequality $\mathbb{E}(X) \leq \int_0^\infty \min\{1, f(t)\} dt$ to yield

$$\begin{aligned} 2 \frac{|\overline{\mathcal{I}}||\overline{\mathcal{J}}|}{mn} \mathbb{E} \left(\sup_{(Q,R,\sigma,\tau) \in \mathcal{Q}} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} r_i g_{\sigma\tau}(\xi_i, \zeta_i) \right| \mid \xi_{\mathcal{I}}, \zeta_{\mathcal{J}} \right) \\ \leq 4 \sqrt{\frac{(|\mathcal{I}| + |\mathcal{J}|) \log K + 2\binom{K}{2} \log(\ell + 1) + \log 2}{2\ell}}. \end{aligned}$$

Since the bound holds for any $\xi_{\mathcal{I}}, \zeta_{\mathcal{J}}$, the same bound holds when the conditioning is removed and $\xi_{\mathcal{I}}, \zeta_{\mathcal{J}}$ are chosen randomly, thus proving the lemma.

B.6. Proof of Lemma 5.7. To abbreviate notation, let $\mathcal{Q} = \mathcal{Q}_V^n \times \mathcal{Q}_U$. Let r_1, \dots, r_m be Rademacher variables as in the proof of Lemma 5.6. By a standard Rademacher symmetrization,

$$\begin{aligned} \mathbb{E} \left(\sup_{(R,\sigma) \in \mathcal{Q}} \left\{ \max_{1 \leq a \leq K} \left| \sigma^{-1}(a) - \frac{1}{m} \sum_{i=1}^m 1\{\sigma(\xi_i) = a\} \right| \right\} \mid \zeta_{\mathcal{J}} \right) \\ \leq 2 \mathbb{E} \left(\sup_{(R,\sigma) \in \mathcal{Q}} \left\{ \max_{1 \leq a \leq K} \left| \frac{1}{m} \sum_{i=1}^m r_i 1\{\sigma(\xi_i) = a\} \right| \right\} \mid \zeta_{\mathcal{J}} \right). \end{aligned}$$

As in the proof of Lemma 5.6, a Hoeffding inequality and union bound yield

$$\begin{aligned} \mathbb{P} \left(\sup_{(R,\sigma) \in \mathcal{Q}} \left\{ \max_{1 \leq a \leq K} \left| \sigma^{-1}(a) - \frac{1}{m} \sum_{i=1}^m 1\{\sigma(\xi_i) = a\} \right| \right\} \geq \epsilon \mid \zeta_{\mathcal{J}} \right) \\ \leq K^{|\mathcal{J}|} (m+1)^{\binom{K}{2}} K \cdot 2e^{-2m\epsilon^2}, \end{aligned}$$

and applying $\mathbb{E}(|X|) \leq \int_0^\infty \min\{1, f(t)\} dt$ for $\mathbb{P}(|X| \geq t) \leq f(t)$ then gives

$$\begin{aligned} 2 \mathbb{E} \left(\sup_{(R, \sigma) \in \mathcal{Q}} \left\{ \max_{1 \leq a \leq K} \left| \sigma^{-1}(a) - \frac{1}{m} \sum_{i=1}^m 1\{\sigma(\xi_i) = a\} \right| \right\} \mid \zeta_{\mathcal{J}} \right) \\ \leq 4 \sqrt{\frac{(|\mathcal{J}| + 1) \log K + \binom{K}{2} \log(m+1) + \log 2}{2m}}. \end{aligned}$$

As in the proof of Lemma 5.6, removing the conditioning on $\zeta_{\mathcal{J}}$ does not alter the bound. Parallel arguments apply to $\tau \in \mathcal{Q}_V$, and the lemma follows.

B.7. Covering argument to establish Theorem 4.1. The establishment of Theorem 4.1 from Proposition 5.1 proceeds as follows. For $\mathcal{F} \subset [0, 1]^{K \times K}$, recall that $h_{\mathcal{F}}(\Gamma) = \sup_{F \in \mathcal{F}} \langle \Gamma, F \rangle = \sup_{F \in \mathcal{F}} \text{tr}(\Gamma^T F)$. By the Cauchy–Schwartz inequality, $h_{\mathcal{F}}$ is Lipschitz continuous:

$$|h_{\mathcal{F}}(\Gamma) - h_{\mathcal{F}}(\Gamma')| \leq \sup_{F \in \mathcal{F}} |\langle \Gamma - \Gamma', F \rangle| \leq K \|\Gamma - \Gamma'\|.$$

Let \mathcal{B}_ϵ denote an ϵ -cover in $\|\cdot\|$ for $[-1, 1]^{K \times K}$, with $\Gamma^{\mathcal{B}}$ the closest point in \mathcal{B}_ϵ to a given Γ . The triangle inequality, Lipschitz condition, and \mathcal{B}_ϵ imply

$$\begin{aligned} \sup_{\Gamma \in [-1, 1]^{K \times K}} |h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)| &\leq \sup_{\Gamma \in [-1, 1]^{K \times K}} \left\{ |h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^A}(\Gamma^{\mathcal{B}})| \right. \\ &\quad \left. + |h_{\mathcal{F}_{\mu\nu}^A}(\Gamma^{\mathcal{B}}) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma^{\mathcal{B}})| + |h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma^{\mathcal{B}}) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)| \right\} \\ &\leq \sup_{\Gamma \in [-1, 1]^{K \times K}} \left\{ |h_{\mathcal{F}_{\mu\nu}^A}(\Gamma^{\mathcal{B}}) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma^{\mathcal{B}})| + 2K \|\Gamma - \Gamma^{\mathcal{B}}\| \right\} \\ &\leq \sup_{\Gamma \in [-1, 1]^{K \times K}} |h_{\mathcal{F}_{\mu\nu}^A}(\Gamma^{\mathcal{B}}) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma^{\mathcal{B}})| + 2K\epsilon \\ &= \max_{\Gamma \in \mathcal{B}_\epsilon} |h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)| + 2K\epsilon. \end{aligned}$$

Now let C_K and n_K be defined as in Proposition 5.1, and set $\epsilon = C_K/n^{1/4}$. It follows by the above relation, a union bound, and Proposition 5.1 that

$$\begin{aligned} \mathbb{P} \left(\max_{(\mu, \nu) \in \Omega_{\rho n} \times \Omega_n} \left\{ \sup_{\Gamma \in [-1, 1]^{K \times K}} |h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)| \right\} \geq 3\epsilon \right) \\ \leq \mathbb{P} \left(\max_{(\mu, \nu) \in \Omega_{\rho n} \times \Omega_n} \left\{ \max_{\Gamma \in \mathcal{B}_{\epsilon/K}} |h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)| \right\} \geq \epsilon \right) \\ \leq \sum_{(\mu, \nu) \in \Omega_{\rho n} \times \Omega_n} \sum_{\Gamma \in \mathcal{B}_{\epsilon/K}} \mathbb{P} \left(|h_{\mathcal{F}_{\mu\nu}^A}(\Gamma) - h_{\mathcal{F}_{\mu\nu}^\omega}(\Gamma)| \geq \epsilon \right) \\ \leq |\Omega_{\rho n}| |\Omega_n| |\mathcal{B}_{\epsilon/K}| 2e^{-\sqrt{n}[2\rho/(\rho+1)]} [1 + o(1)] \end{aligned}$$

for all $n \geq n_K$. The result of Theorem 4.1 then follows, since we have that $|\Omega_n| = \binom{n+K-1}{K-1}$, and $\mathcal{B}_{\epsilon/K}$ can be chosen such that $|\mathcal{B}_{\epsilon/K}| \leq (1 + K^2/\epsilon)^{K^2}$.

ACKNOWLEDGEMENTS

The first author wishes to thank Peter Bickel for helpful advice and feedback. Work supported in part by the US Army Research Office under PECASE Award W911NF-09-1-0555, the US Office of Naval Research under MURI Award 58153-MA-MUR, the UK Engineering and Physical Sciences Research Council under an Institutional Sponsorship Award, and the UK Royal Society under a Wolfson Research Merit Award.

REFERENCES

- [1] AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.
- [2] ALON, N., DE LA VEGA, W. F., KANNAN, R. and KARPINSKI, M. (2003). Random sampling and approximation of MAX-CSPs. *J. Comput. System Sci.* **67** 212–243.
- [3] BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.
- [4] BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39** 2280–2301.
- [5] BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T., SZEGEDY, B. and VESZTERGOMBI, K. (2006). Graph limits and parameter testing. In *Proc. ACM Symp. Theory Comput.* 261–270.
- [6] BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T. and VESZTERGOMBI, K. (2008). Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing. *Adv. Math.* **219** 1801–1851.
- [7] BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T. and VESZTERGOMBI, K. (2012). Convergent sequences of dense graphs II. Multiway cuts and statistical physics. *Ann. Math.* **176** 151–219.
- [8] BOUSQUET, O., BOUCHERON, S. and LUGOSI, G. (2004). Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning* (O. Bousquet, U. von Luxburg and G. Rätsch, eds.) 169–207. Springer, Berlin.
- [9] CHATTERJEE, S. (2012). Matrix estimation by universal singular value thresholding. Preprint arXiv:1212.1247.
- [10] CHOI, D. S., WOLFE, P. J. and AIROLDI, E. O. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284.
- [11] CLEMENÇON, S., LUGOSI, G. and VAYATIS, N. (2008). Ranking and empirical minimization of U-statistics. *Ann. Statist.* **36** 844–874.
- [12] DIACONIS, P. and JANSON, S. (2008). Graph limits and exchangeable random graphs. *Rendiconti di Matematica* **28** 33–61.
- [13] FIENBERG, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *J. Computat. Graph. Statist.* **21** 825–839.
- [14] FISHKIND, D. E., SUSSMAN, D. L., TANG, M., VOGELSTEIN, J. T. and PRIEBE, C. E. (2012). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. Preprint arXiv:1205.0309.

- [15] FLYNN, C. J. and PERRY, P. O. (2012). Consistent biclustering. Preprint arXiv:1206.6927.
- [16] FORTUNATO, S. and BARTHELEMY, M. (2007). Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA* **104** 36–41.
- [17] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.
- [18] HOFF, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Computat. Math. Org. Theory* **15** 261–272.
- [19] HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098.
- [20] KIM, M. and LESKOVEC, J. (2012). Multiplicative attribute graph model of real-world networks. *Internet Math.* **8** 113–160.
- [21] MILLER, K. T., GRIFFITHS, T. L. and JORDAN, M. I. (2009). Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems 22* (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta, eds.) 1276–1284.
- [22] NEWMAN, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8582.
- [23] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915.
- [24] ROHE, K. and YU, B. (2012). Co-clustering for directed graphs: The stochastic co-blockmodel and a spectral algorithm. Preprint arXiv:1204.2296.
- [25] SCHNEIDER, R. (1993). *Convex Bodies: The Brunn–Minkowski Theory*. Cambridge University Press, Cambridge, UK.
- [26] ZHAO, Y., LEVINA, E. and ZHU, J. (2011). Community extraction for social networks. *Proc. Natl. Acad. Sci. USA* **108** 7321–7326.
- [27] ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* To appear.

HEINZ COLLEGE OF PUBLIC POLICY AND MANAGEMENT
 CARNEGIE MELLON UNIVERSITY
 5000 FORBES AVE, HAMBURG HALL
 PITTSBURGH, PA 15213-3890, USA
 E-MAIL: davidch@andrew.cmu.edu

DEPARTMENT OF STATISTICAL SCIENCE
 UNIVERSITY COLLEGE LONDON
 GOWER STREET
 LONDON WC1E 6BT, UK
 E-MAIL: patrick@stats.ucl.ac.uk